

The genome of the medieval Black Death agent extended abstract

Ashok Rajaraman^{1,2}, Eric Tannier^{3,4}, Cedric Chauve^{1,5}

¹ Department of Mathematics, Simon Fraser University, V5A 1S6 Burnaby (BC), Canada
{arajaram, cedric.chauve}@sfu.ca

² International Graduate Training Center in Mathematical Biology, Pacific Institute for Mathematical Sciences,
Vancouver (BC), Canada

³ INRIA Grenoble Rhône-Alpes, F-38334 Montbonnot, France
eric.tannier@inria.fr

⁴ Université de Lyon 1, Laboratoire de Biométrie et Biologie Évolutive, CNRS UMR5558 F-69622 Villeurbanne, France

⁵ LaBRI, Université Bordeaux I, 33405 Talence, France

Abstract *The genome of a 650 year old Yersinia pestis bacteria, responsible for the medieval Black Death, was recently sequenced and assembled into 2,105 contigs from the main chromosome. According to the point mutation record, the medieval bacteria could be an ancestor of most Yersinia pestis extant species, which opens the way to reconstructing the organization of these contigs using a comparative approach. We show that recent computational paleogenomics methods, aiming at reconstructing the organization of ancestral genomes from the comparison of extant genomes, can be used to correct, order and complete the contig set of the Black Death agent genome, providing a full chromosome sequence, at the nucleotide scale, of this ancient bacteria. This sequence suggests that a burst of mobile elements insertions predated the Black Death, leading to an exceptional genome plasticity and increase in rearrangement rate.*

Keywords Paleogenomics, computational biology, genome assembly, pathogens

Le génome de la bactérie responsable de la Peste Noire

Résumé *Récemment, le génome d'une souche de la bactérie Yersinia pestis vieille de 650 ans a été séquencée et assemblée en 2,105 contigs issus de son chromosome. Cette bactérie médiévale semble être l'ancêtre de la plupart des souches actuelles de Yersinia pestis, ce qui permet d'appliquer une approche comparative pour assembler ces contigs en scaffolds. En utilisant des méthodes et principes récemment développés pour la reconstruction de l'organisation de génomes anciens à partir de la comparaison de génomes existants, nous corrigeons, organisons et complétons les contigs de l'agent de la Peste Noire, pour obtenir une séquence complète pour le chromosome de cette bactérie ancienne. L'analyse de cette séquence suggère que de nombreuses insertions d'éléments mobiles ont participé à l'émergence d'un génome exceptionnellement dynamique et à une augmentation du taux de réarrangements.*

Mots-clés Paléogénomique, bioinformatique, assemblage de génomes, pathogènes.

1 Introduction

The plague has long been among the most feared human diseases [10], due to dramatic pandemics such as the *Black Death* which ravaged Europe in the late middle-ages. Recently Bos *et al.* [8] were able to sequence the whole genome of the Black Death agent, and concluded that it was an ancestor of most extant strains of the human pathogen *Yersinia pestis* (see also [44]). The sequence extracted from the oral metagenome of one individual was assembled using Velvet [48], into approximately 130,000 contigs, including 2,105 contigs of length ≥ 500 bp from the main chromosome, with similarities with some *Yersinia* extant genomes¹. This

1. There are 2134 provided contigs in total, and we discarded the 29 ones with no similarities with any *Yersinia* extant genome because they are likely to be artefactual.

first sequencing of the chromosome of an extinct prokaryote helped to understand the causes of the Black Death pandemic [8,37,47]. However, the assembled 2,105 contigs cover only 85% of the expected length of the ancestral chromosome and their organization along this ancestral chromosome is unknown, keeping out of reach a detailed genome-scale study of the evolution of the structural organization of *Yersinia* genomes, whose impact on pathogenicity is still an important open question [11].

Current assembly methodologies can hardly be applied to fully assemble and finish an ancient genome, aside of short molecules such as plasmids [44] and organelle genomes [36]. Indeed, existing scaffolding methods, aimed at ordering and orienting the contigs, and estimating the lengths of inter-contig gaps, rely on additional data such as mate-pair reads with mixed insert sizes [2,40,22,41,49], optical or physical maps [27] or comparison with one or several closely related genomes [42,24]. However, due to the decay and fragmentation of ancient DNA, reads from ancient genomes are in general short, and optical maps or mate-pair libraries with long inserts can not be obtained. This leaves the comparative approach as the only possibility to scaffold large ancient genomes. The usual setting of the comparative approach involves the comparison of the contigs with one, or a few, closely related genomes, either genome sequence or maps [7,5,34,24] or protein sequences [42]. However, to the best of our knowledge, none of these methods is intended to be applied on the genome of an internal node of a given phylogeny.

We describe a comparative approach to scaffold an ancient genome, and apply it to the medieval plague agent. The ancestral Black Death agent is indeed related to a dozen of descendants (from the *Yersinia pestis* clade) and close outgroups (from the *Yersinia pestis* and *Yersinia pseudotuberculosis* clades), whose phylogeny, taken from Bos *et al.* [8], is shown on Fig. 1.

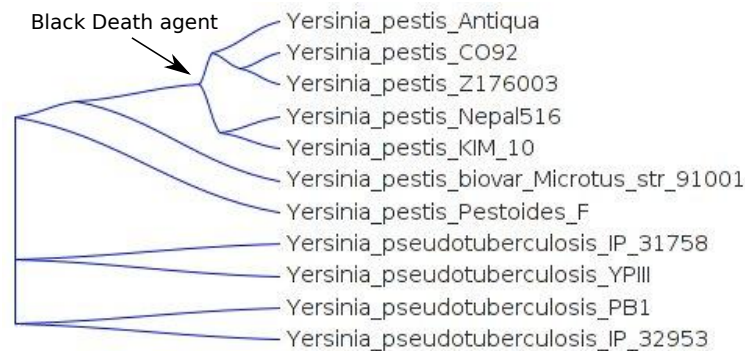


Figure 1. Phylogeny of the used extant genomes and position of the reconstructed one.

There has been a recent flurry of ancestral genome organization reconstruction methods, complementing classical methods for reconstructing ancestral genome sequence [6,26,19] and gene content [15,16,45]. They have been used for reconstructing ancestral genomes of bacteria [46,21], animals [9,29,39,33,38,14,1,31,35], plants [43,32], yeasts [23,13,4] or protists [28]. Recent developments provide exact and fast algorithms that handle sequence duplications, repeats, diverse types of genome rearrangements and chromosome structures [3,25,30].

We show here that this corpus of methods is efficient and versatile enough to be integrated into a comparative scaffolding framework for ancient bacterial genomes, and we illustrate this claim with a complete assembly of the medieval Black Death agent chromosome. Starting from the contigs assembled by Bos *et al.* [8] which have similarities with extant *Yersinia* genomes, we compute a single circular scaffold containing the ordered and oriented sequences from the whole set of contigs, completed by estimations of the sequences located between consecutive contigs (gaps). Additionally, we correct some contigs initially assembled by Bos *et al.* by identifying probable *chimeric*, *redundant* or *duplicated* contigs. The chromosome structure we observe is distant from every extant genome, explaining the difficulty of the assembly process with a single reference genome. We annotate and analyse the ancestral chromosome, pointing at a probable replication origin, predicting the positions of insertion sequences (IS) and detecting the numerous inversions that separate it from extant genomes. We provide evidence that the speciation between the *Yersinia pestis* and *Yersinia pseudotuberculosis*

clades was characterized by a burst of insertion of IS elements in the *Yersinia pestis* genomes, concomitant with an increase rate of genome rearrangements, which breakpoints positions are also correlated with IS.

2 Results

The main result of our work is a completely assembled chromosome sequence of the Black Death agent genome. To obtain it, we followed a generic procedure for reconstructing an ancestral genome organization [29,14,31,4,25], which comprises four phases: (1) extracting homologous families of ancestral and extant genome markers, (2) computing putative linkage between ancestral markers, (3) combining the set of ancestral linkages into a circular sequence of ancestral markers, (4) inferring inter-marker gap sequences. We provide only a sketch of the implementation in this extended abstract, and full details will be published elsewhere.

Families of homologous segments. We aligned the ancestral contigs against 11 fully assembled genomes of *Yersinia* strains. Several contigs were not aligned over their full length on every genome because of rearrangements. So we cut the contigs into pieces, such that every piece is aligned over its full length on every genome and no pair of genomic segment defined by two different alignments overlap (they are either disjoint or confounded). This clusters ancestral and extant genome segments into 2,619 homologous families. Each family contains one or several ancestral contig segments, and zero, one or several genome segments from each extant species.

All sequences from a single family are assumed to be homologous, that is, they share a common ancestor and having evolved through speciations, duplications, losses or transfers. We do not have phylogenetic trees for the families that would allow us to detect those events and derive a marker content [45]. Yet some ancestral markers correspond to repeated sequences that were present at several loci of the ancestral genome, while some of them contain ancestral segments from several different contigs. We used phyletic profiles [15,16] to determine the number of occurrences of every ancestral marker, namely the ancestral marker content of this ancestral genome. We computed this ancestral content for each family by using a parsimony approach that minimizes the number of gains and losses of markers along the species tree for each family. This allows to associate to each family a *multiplicity*, *i.e.* its expected number of occurrences in the ancestral chromosome; 20 families out of 2,619 have a multiplicity greater than 1.

The amount of DNA encoded by the markers, when multiplicity is accounted for, is 3,846,866bp of ancestral DNA, while the initial contigs encode 4,013,159bp. This initial loss of sequenced ancient DNA will be compensated by filling the gaps between the different pieces of the segmented contigs.

Computing putative linkages between ancestral markers. We computed sets of ancestral markers that are believed to be consecutive in the ancestral chromosome. We call them *intervals* of ancestral markers, if they contain more than two markers and *adjacencies* if they concern only two markers. We followed a Dollo parsimony principle [14] to infer putative ancestral linkages: a group of ancestral markers is deemed to be contiguous in the ancestral genome if markers from the same families are contiguous in at least two extant genomes whose evolutionary path on the species phylogeny contains the ancestor of interest (here the Black Death agent). All 2,637 putative adjacencies obtained in this way are then weighted according to their phylogenetic conservation, using a recursive formula inspired from the Fitch-Hartigan principle [29,14,4].

Combining the set of ancestral linkages into a circular sequence of ancestral markers. The set of putative ancestral adjacencies is not compatible with a circular chromosomal structure, due to possible converging genome rearrangements, for example. Indeed some markers may be involved in too many adjacencies. However, discarding 6 adjacencies out of the 2,637 putative ancestral adjacencies was enough to obtain a set of maximal cumulative weight that can be ordered circularly. They were found implementing a fast and exact "circularization" method based on matching techniques in graphs[30].

Adjacencies alone are compatible with many circular orders due to repeated ancestral markers forming tangles in the adjacency graph [24,2]. To address this issue, intervals of size greater than two were used as illustrated in Fig. 2 to clear the ambiguities, resulting in an ordering of the markers into three large scaffolds.

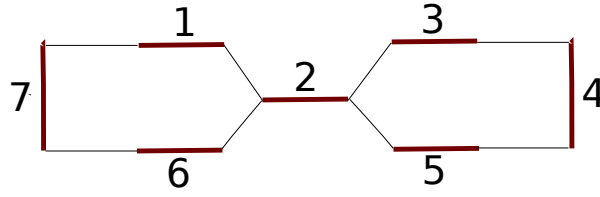


Figure 2. Illustration of the ambiguity in ordering ancestral markers with multiplicities greater than 1 and of the use of intervals to address it. Here is a toy example where we have markers $1, \dots, 7$, drawn with bold red segments, and adjacencies between their extremities, drawn with thin black lines. Assume every marker has multiplicity 1 except marker 2, which has multiplicity 2. Then every marker extremity has as many adjacencies as its multiplicity predicts. But there are several possible circular orderings of these markers according to these adjacencies: $1, 2, 3, 4, 5, 2, 6, 7$, or $1, 2, 5, 4, 3, 2, 6, 7$. Suppose we have in addition size three intervals, and among them we find $\{1, 2, 3\}$ or $\{2, 5, 6\}$. Then only the first ordering is compatible. In our data set, intervals up to size 6 were sufficient to completely clarify the adjacency signal.

We then joined the extremities of these three scaffolds to form a circular chromosome by choosing, among the six possible configurations, the only one supported by some extant genomes. This resulted into a complete circular ordering of ancestral markers, where each ancestral marker appears exactly as many times as it is expected from its multiplicity.

Correcting the initial contigs. In the resulting ordering, each occurrence of an ancestral marker corresponds to one or several segments of the initial contigs. The ordering of these segments is mostly compatible with the initial contigs. We found only one *chimeric* contig (see Fig. 3), split into two non-adjacent markers in the ancestral genome organization. None of the extant occurrences from the two families are adjacent in extant genomes, pointing to either an assembly error during the initial contig construction, or a derived rearrangement in the ancient genome, which would be interesting since Bos *et al* [8] did not find such a mutation looking at nucleotide substitutions. Note that the length filtering applied onto families after the contig segmentation step can lead to an underestimation of the number of chimeric contigs: if part of a contig has length less than the threshold, it is discarded and the contig is not detected as chimeric. Also four contigs segments were found to be *duplicated*: a large part ($> 500\text{bp}$) of each is probably present in more than one occurrence in the ancestral genome, while the initial assembly predicted only one occurrence. Finally, 63 contigs have a sequence which is found, up to very small variations, inside another contig while their number of extant occurrences suggest they have multiplicity one, so we believe they are *redundant*.

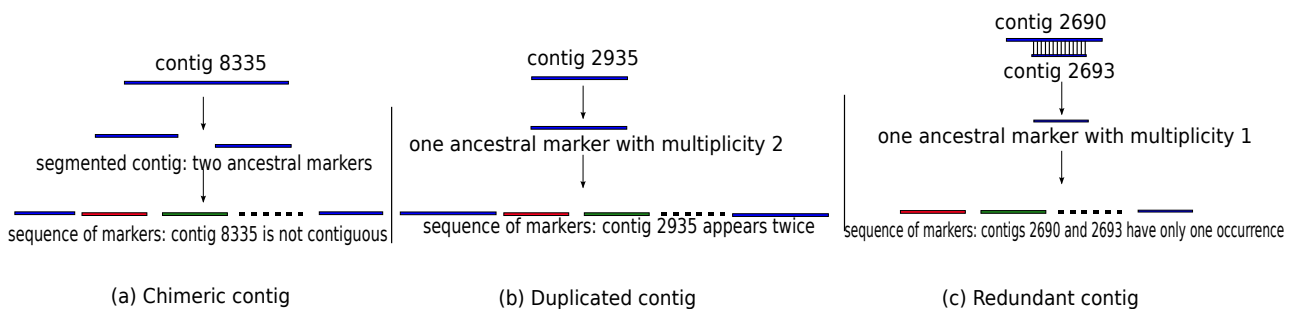


Figure 3. Contig correction: (a) the contig is cut during the segmentation procedure, but not joined during the marker ordering; (b) the contig is found to have two occurrences in the marker ordering; (c) two contigs contain the same DNA sequence and this sequence is predicted to have only one occurrence in the marker ordering.

Estimating ancestral gaps sequences. We completed this assembly by estimating the sequences located in *ancestral gaps*, i.e. between pairs of ancestral markers consecutive in the circular ordering. For this we first

estimated a length interval for each ancestral gap: a length is said to be *supported* for an ancestral gap if there are two gaps in extant genomes, in two species whose evolutionary path contains the ancestor of interest, with such a length. The length interval of a gap is defined by the minimum and maximum supported length for this gap. For 24 gaps we found no supported length, so we took the minimum and maximum gap length of extant sequences in the species where the markers are consecutive. Then for each ancestral gap, we aligned all extant gaps which lengths fall in the ancestral gap length interval. We then constructed an ancestral sequence from each alignment by an ancestral discrete character reconstruction method implementing the Fitch algorithm [20].

This resulted in an ancestral genome sequence of length 4,586,856 showing that 739,990bp were added to the ancestral markers sequences by this finishing step. Only 1 gap was not assigned a sequence by this method.

Analysis of the reconstructed ancestor. We took advantage of reconstructing the full chromosome of the Black Death agent to analyze its structure and evolution at the whole-genome scale.

We traced the GC-skew with SeqinR [12] from a CDS annotation by Glimmer (Fig. 4(b)) predict the position of the replication origin. We slipped the medieval sequence such that the putative replication origin (the maximum value in the cumulative GC-skew plot) has position 0 and we aligned the ancient chromosome with the chromosome of the CO92 strain. We obtained the dotplot represented in Fig. 4(a) that shows the highly repeated nature of both genomes, and the rearrangements that have happened along the lineage from the ancestor to the CO92 strain.

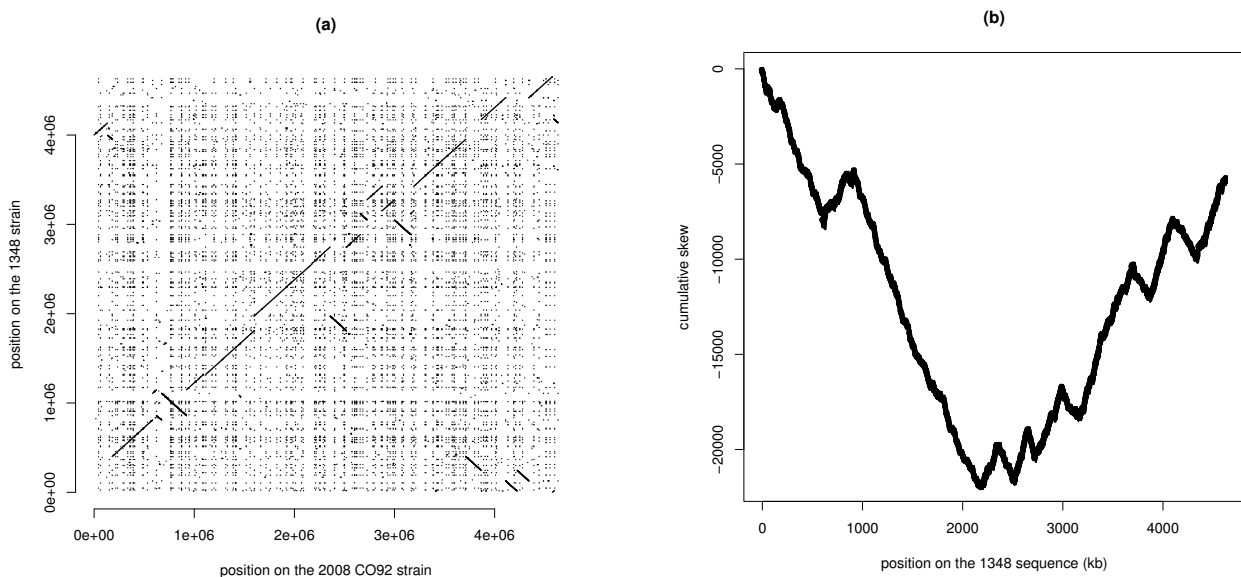


Figure 4. (a) Dotplot of all Megablast alignments of the medieval sequence against the CO92 extant strain. The highly repetitive nature of both genomes appear, as well as the inversions that happened in the CO92 history, several of them being symmetric around the origin of replication. (b) Cumulative skew shows a probable position for the replication origin (for which we chose position 0), as well as the rearrangements which tend to blur the skew signal.

We mapped IS elements onto the reconstructed ancestral chromosome, based on a conservative analysis of their patterns of presence in extant markers and gaps: an ancestral gap is assigned an IS if one of its occurrences in the descendants genomes is of length exactly the minimum length of the ancestral gap and contains an annotated IS; we focused on gaps as no extant marker does contain an annotated IS. This resulted in 94 ancestral gaps containing IS. We confirmed this comparative annotation with an automatic annotation. Our analysis also shows that a large part of these IS (at least 57) were already present in the last common ancestor of all *Yersinia pestis* strains, while they are almost completely absent from the genomes of the considered *Yersinia pseudotuberculosis*.

We also analysed the genome rearrangements between the ancestral sequence and extant genomes by sampling inversion scenarios between the ancestral genome and the extant genomes (see Fig. 5). There are 8-9

inversions between the *Yersinia pseudotuberculosis* strains and the medieval genome, and 9-22 inversions when compared to (thought evolutionarily closer) *Yersinia pestis* strains. As noticed by Darling et al [17], we can also observe that inversion breakpoints are not randomly distributed and used: highly used ones are concentrated in one third of the chromosome, around its replication origin. Most inversions are symmetrical around the origin. The positions of the inversion breakpoints are also highly correlated with IS, as remarked earlier [18]: 76 out of the 118 mapped breakpoints are close ($< 1000\text{bp}$ distant) to some predicted IS, while this number drops to 39 for uniformly sampled random coordinates ($p\text{-value} < 10^{-3}$). Rearrangements are very numerous in all *pestis* branches, strongly suggesting that they could be driven by the IS.

<i>Yersinia pestis</i> biovar Microtus str 91001	22
<i>Yersinia pestis</i> Pestoides F	13
<i>Yersinia pseudotuberculosis</i> IP 31758	9
<i>Yersinia pseudotuberculosis</i> YPIII	8
<i>Yersinia pseudotuberculosis</i> PB1	9
<i>Yersinia pseudotuberculosis</i> IP 32953	8
<i>Yersinia pestis</i> Antiqua	21-22
<i>Yersinia pestis</i> CO92	12
<i>Yersinia pestis</i> Z176003	13
<i>Yersinia pestis</i> Nepal516	9
<i>Yersinia pestis</i> KIM 10	9

Figure 5. Rearrangement distances between the extinct genome and the extant genomes. Two numbers mean that sampled scenarios have different length as we sample scenarios following a Bayesian posterior distribution of all scenarios, and not only the most parsimonious ones.

3 Discussion/Conclusion

The present work illustrates the potential of phylogenetic/comparative assembly methods to address the specific issues of ancient DNA assembly (single reads, fragmentation, ...). Our main result is a complete assembly of the chromosome of a 650 years old bacteria, that opens the way to whole genome analysis of rearrangements and insertion dynamics among others.

The method we developed for this assembly relies on recent advances, both methodological and algorithmic, in reconstructing the organization of ancient genomes from the comparison of related extant genomes. We show here that such methods are generic enough to be also used with data acquired by sequencing of ancient DNA.

A crucial issue of such a method is its validation. In this extended abstract we do not develop this point but we are currently extensively testing our method on simulated data generated from *Yersinia* genomes.

We believe the methodological advances we present in this work complement the recent breakthrough in ancient DNA sequencing, at least for bacterial genomes, and suggest that integrating ancient genomes into comparative genomics is an ambitious but realistic goal for the next few years.

Acknowledgements

This work was supported by NSERC Discovery Grant to C.C., a PIMS IGTC Fellowship to A.R. and ANR-10-BINF-01-01 Ancestrum to E.T. We are thankful to Laurent Duret, Vincent Daubin, Annie Chateau, Eric Rivals, Hendrik Poinar for useful discussions.

References

- [1] M. A. Alekseyev, P. A. Pevzner. Breakpoint graphs and ancestral genome reconstructions. *Genome Res*, 19(5):943–957, 2009.
- [2] A. Bashir, A. Klammer, W. P. Robins, et al. A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotech*, 30(7):701–707, 2012.

- [3] S. Bérard, C. Gallien, B. Boussau, G. J. Szollosi, Vincent Daubin, Eric Tannier. Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics*, 28:i382–i388, 2012.
- [4] D. Bertrand, Y. Gagnon, M. Blanchette, N. El-Mabrouk. Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss. In Mona Singh and Vincent Moulton, editors, *Algorithms in Bioinformatics, 10th International Workshop, WABI 2010, Liverpool, UK, September 6-8, 2010. Proceedings*, volume 6293 of *Lecture Notes in Bioinformatics*, pages 78– 89. Springer Verlag, 2010.
- [5] D. Bertrand, M. Blanchette, N. El-Mabrouk. Genetic map refinement using a comparative genomic approach. *J Comput Biol*, 16(10):1475–1486, 2009.
- [6] M. Blanchette, E. D. Green, W. Miller, D. Haussler. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*, 14(12):2412–2423, 2004.
- [7] G. Blin, E. Blais, D. Hermelin, P. Guillon, M. Blanchette, N. El-Mabrouk. Gene maps linearization using genomic rearrangement distances. *J Comput Biol*, 14(4):394–407, 2007.
- [8] K. I. Bos, V. J. Schuenemann, G. B. Golding, et al. A draft genome of yersinia pestis from victims of the black death. *Nature*, 478(7370):506–510, 2011.
- [9] G. Bourque, P. A. Pevzner, G. Tesler. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res*, 14(4):507–516, 2004.
- [10] A. Camus. *La peste*. Gallimard, 1947.
- [11] P.S. Chain, E. Carniel, F.W. Larimer, et al. Insights into the evolution of yersinia pestis through whole-genome comparison with yersinia pseudotuberculosis. *Proc Natl Acad Sci U S A*, 101(38):13826–13831, 2004.
- [12] D. Charif, J.R. Lobry. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, editors, *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York, 2007. ISBN : 978-3-540-35305-8.
- [13] C. Chauve, H. Gavranovic, A. Ouangraoua, E. Tannier. Yeast ancestral genome reconstructions: the possibilities of computational methods II. *J Comput Biol*, 17(9):1097–1112, 2010.
- [14] C. Chauve, E. Tannier. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput Biol*, 4(11):e1000234, 2008.
- [15] O. Cohen, H. Ashkenazy, F. Belinky, D. Huchon, T. Pupko. Gloome: gain loss mapping engine. *Bioinformatics*, 26(22):2914–2915, 2010.
- [16] M. Csurös. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912, 2010.
- [17] A. E. Darling, I. Miklós, M. A. Ragan. Dynamics of genome rearrangement in bacterial populations. *PLoS Genet*, 4(7):e1000128, 2008.
- [18] W. Deng, V. Burland, G. Plunkett, et al. Genome sequence of yersinia pestis kim. *J Bacteriol*, 184(16):4601–4611, 2002.
- [19] A. B. Diallo, V. Makarenkov, M. Blanchette. Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics*, 26(1):130–131, 2010.
- [20] W. M. Fitch. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst Zool*, 20(4):406–416, 1971.
- [21] R. Frenez, T. Faraut, G. Fichant, J. Gouzy, Y. Quentin. Phylogenetic exploration of bacterial genomic rearrangements. *Bioinformatics*, 23(9):1172–1174, 2007.
- [22] S. Gao, W.-K. Sung, N. Nagarajan. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *J Comput Biol*, 18:1681–1691, 2011.
- [23] J. L. Gordon, K. P. Byrne, K. H. Wolfe. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern saccharomyces cerevisiae genome. *PLoS Genet*, 5(5):e1000485, 2009.
- [24] P. Husemann, J. Stoye. Phylogenetic comparative assembly. *Algorithms Mol Biol*, 5:3, 2010.
- [25] B. R. Jones, A. Rajaraman, E. Tannier, C. Chauve. ANGES: Reconstructing ancestral genomes maps. *Bioinformatics*, 28:2388–2390, 2012.
- [26] D. A. Liberles, editor. *Ancestral Sequence Reconstruction*. Oxford University Press, 2007.
- [27] H. C. Lin, S. Goldstein, L. Mendelowitz, S. Zhou, J. Wetzel, D. C. Schwartz, M. Pop. Agora: Assembly guided by optical restriction alignment. *BMC bioinformatics*, 13:189, 2012.

- [28] J. Ma, A. Ratan, B. J. Shuh, L. Zhang, W. Miller, D. Haussler. Dupcar: reconstructing contiguous ancestral regions with duplications. *J Comput Biol*, 15:1007–1027, 2008.
- [29] J. Ma, L. Zhang, B. B. Suh, B. J. Raney, R. C. Burhans, W. J. Kent, M. Blanchette, D. Haussler, W. Miller. Reconstructing contiguous regions of an ancestral genome. *Genome Res*, 16(12):1557–1565, 2006.
- [30] J. Mañuch, M. Patterson, R. Wittler, C. Chauve, E. Tannier. Linearization of ancestral multichromosomal genomes. *BMC Bioinformatics*, 13(Suppl 19):S11, 2012.
- [31] M. Muffato, A. Louis, C.-E. Poisnel, H. Roest Crolius. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, 26(8):1119–1121, 2010.
- [32] F. Murat, J.-H. Xu, E. Tannier, M. Abrouk, N. Guilhot, C. Pont, J. Messing, J. Salse. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res*, 20(11):1545–1557, 2010.
- [33] Y. Nakatani, H. Takeda, Y. Kohara, S. Morishita. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res*, 17(9):1254–1265, 2007.
- [34] A. Muñoz, C. Zheng, Q. Zhu, V. A. Albert, S. Rounsley, D. Sankoff. Scaffold filling, contig fusion and comparative gene order inference. *BMC Bioinformatics*, 11:304, 2010.
- [35] A. Ouangraoua, E. Tannier, C. Chauve. Reconstructing the architecture of the ancestral amniote genome. *Bioinformatics*, 27(19):2664–2671, 2011.
- [36] J. L. Paijmans, M. T. Gilbert, M. Hofreiter. Mitogenomic analyses from ancient dna. *Mol Phylogenet Evol*, 1012. Epub ahead of print (Jun 15, 2012).
- [37] J. Parkhill, B. W. Wren. Bacterial epidemiology and biology - lessons from genome sequencing. *Genome Biol*, 12:230, 2011.
- [38] N. H. Putnam, T. Butts, D. E. K. Ferrier, et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198):1064–1071, 2008.
- [39] N. H. Putnam, M. Srivastava, U. Hellsten, et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, 317(5834):86–94, 2007.
- [40] F. J. Ribeiro, D. Przybylski, S. Yin, et al. Finished bacterial genomes from shotgun sequence data. *Genome Res*, 22:2270–2277, 2012.
- [41] L. Salmela, V. Mäkinen, N. Välimäki, J. Ylinen, E. Ukkonen. Fast scaffolding with small independent mixed integer programs. *Bioinformatics*, 27:3259–3265, 2011.
- [42] S. L. Salzberg, D. D. Sommer, D. Puiu, V. T. Lee. Gene-boosted assembly of a novel bacterial genome from very short reads. *PLoS Comput Biol*, 4:e1000186, 2008.
- [43] D. Sankoff, C. Zheng, P. K. Wall, C. dePamphilis, J. Leebens-Mack, V. A. Albert. Towards improved reconstruction of ancestral gene order in angiosperm phylogeny. *J Comput Biol*, 16(10):1353–1367, 2009.
- [44] V. J. Schuenemann, K. I. Bos, S. DeWitte, et al. Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of yersinia pestis from victims of the black death. *Proc Natl Acad Sci U S A*, 108:E746–E752, 2011.
- [45] G. J. Szöllősi, B. Boussau, S. S. Abby, E. Tannier, V. Daubin. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci U S A*, 109:17513–17518, 2012.
- [46] Y. Wang, W. Li, T. Zhang, C. Ding, Z. Lu, N. Long, J. P. Rose, B.-C. Wang, D. Lin. Reconstruction of ancient genome and gene order from complete microbial genome sequences. *J Theoret Biol*, 239:494–498, 2006.
- [47] D. J. Wilson. Insights from genomics into bacterial pathogen populations. *PLoS Pathog*, 8:e1002874, 2012.
- [48] D. R. Zerbino, E. Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res*, 18(5):821–829, 2008.
- [49] D. R. Zerbino, G. K. McEwen, E. H. Margulies, E. Birney. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One*, 4:e8407, 2009.